D408–D411 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 21 October 2008*
doi:10.1093/nar/gkn749

PEDANT covers all complete RefSeq genomes

Mathias C. Walter[1], Thomas Rattei[2], Roland Arnold[2], Ulrich Güldener[1], Martin Münsterkötter[1], Karamfilka Nenova[1], Gabi Kastenmüller[1], Patrick Tischler[2], Andreas Wölling[3], Andreas Volz[3], Norbert Pongratz[3], Ralf Jost[3], Hans-Werner Mewes[1,2] and Dmitrij Frishman[1,2,*]

[1]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Ingolstädter Landstrasse 1, 85764 Neuherberg, [2]Department of Genome-oriented Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, Am Forum 1, 85350 Freising and [3]Biomax Informatics AG, Lochhamer Strasse 9, 82152 Martinsried, Germany

Received September 15, 2008; Accepted October 3, 2008

ABSTRACT

The PEDANT genome database provides exhaustive annotation of nearly 3000 publicly available eukaryotic, eubacterial, archaeal and viral genomes with more than 4.5 million proteins by a broad set of bioinformatics algorithms. In particular, all completely sequenced genomes from the NCBI's Reference Sequence collection (RefSeq) are covered. The PEDANT processing pipeline has been sped up by an order of magnitude through the utilization of precalculated similarity information stored in the similarity matrix of proteins (SIMAP) database, making it possible to process newly sequenced genomes immediately as they become available. PEDANT is freely accessible to academic users at http://pedant.gsf.de. For programmatic access Web Services are available at http://pedant.gsf.de/webservices.jsp.

INTRODUCTION

Since its first announcement in 1997 (1), the PEDANT genome database has steadily grown to become one of the most comprehensive collections of automatically annotated genomes. As of September 2008, PEDANT covers all complete genomes as provided by the RefSeq (2) database. In total 861 completely sequenced genomes from all three domains of life as well as 2081 complete viral genomes are available (Table 1). Here, we define a 'complete genome' as a genome whose chromosomal datasets exist as RefSeq records or Ensembl (3) entries and genes have been predicted. For those eukaryotic genomes (currently 33) that are available both from RefSeq or Ensembl, we provide the annotation of both versions. This results in a total number of 2975 genome databases with 4.5 million proteins occupying 3.1 TB of storage. All PEDANT databases are continuously updated. For example, assignments of genes to the MIPS Functional Catalog (FunCat) (4) have been recently recalculated using the new 2.1 version of FunCat (http://mips.gsf.de/projects/funcat).

The current version of the software driving the PEDANT web site, which we refer to as PEDANT3, represents an industry-strength Java workbench that supports large-scale grid computing and utilizes a work-flow-based processing engine (D. Frishman *et al.*, manuscript in preparation). Dozens of custom workflows are available: generic workflows for eukaryotic, prokaryotic and viral genomes as well as more specialized workflows supporting specific genome groups (gram-positive versus gram-negative bacteria, fungi, plants), data types (EST collections, raw contigs without any predicted Open Reading Frames (ORFs), protein-only datasets, etc.) and bioinformatics methods (e.g. alternative gene prediction techniques). Advanced protein and DNA viewers implemented using server-side Java provide graphical representation of protein annotation features as well as genetic elements on chromosomes.

NEW FEATURES AND IMPROVEMENTS

Genome import pipeline

Given the quick pace of genome sequencing keeping track of currently available data and obtaining them from source databases for local processing represents a time-consuming and technically challenging task. In order to organize a more efficient import of genomes to PEDANT from various sources, we set up a specialized processing pipeline (Figure 1). In the first step, we acquire a list of available genomes from each genome resource. Then we

*To whom correspondence should be addressed. Tel: +49 8161 712134; Fax: +49 8161 712186; Email: d.frishman@wzw.tum.de

Table 1. The number of species from major taxonomic groups contained in the PEDANT genome database as of September 2008

NCBI Taxonomy ID	Taxonomic group	Number of genomes
131567	Cellular organisms	861
2157	Archaea	53
2	Bacteria	691
2759	Eukaryota	117
4751	Fungi	*57*
33208	Metazoa	*42*
33090	Viridiplantae	*6*
	Other	*12*
10239	Viruses	*2081*
	Total	2942

Other groups: Alveolata (2), Amoebozoa (1), Cryptophyta (1), Euglenozoa (5), Rhodophyta (1), Stramenopiles (2).

try to find out the Entrez genome project ID by using the Entrez Programming Utilities (eUtils, http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html) and querying the NCBI databases (5) for genome project information. If available, we use the genome project ID as a primary key for a given genome, otherwise the NCBI taxonomy ID is utilized. The advantage of genome project IDs is that they are stable in contrast to the taxonomy IDs which may change, especially for the species/strains of newly sequenced genomes. The genome IDs are then stored in our local meta-database which also serves as the data basis for generating the full genome list for the PEDANT web page.

Data retrieval procedures have been adapted to several different sources of genome information. For downloading RefSeq genomes, we use a patched version (retry on connection timeouts, improved error handling) of the NCBI ToolBox (http://www.ncbi.nlm.nih.gov/IEB/ToolBox) program. For Ensembl genomes, we install the provided MySQL database dumps (ftp://ftp.ensembl.org/pub/current_mysql) at our local MySQL server and extract the genomic data directly.

Retrieval of genomes not contained in RefSeq and Ensembl can only be done in a semi-automatic fashion with manual verification. In many cases, RefSeq lists the involved genome sequence centers where original data can be obtained. Another useful resource to locate genomes is 'the genomes online database (GOLD)' (6). We then retrieve the assembly and annotation data directly from the sequence centers and check them for missing sequences, nonunique identifiers and unusual formatting. If the gene annotation data are missing or in a draft version (especially fungal genomes), gene predictions are carried out or existing models are improved dependent on the annotation project (7,8).

Integration of PEDANT and SIMAP

Calculating and updating protein similarities and domain assignments is the most time consuming and computationally expensive task in our genome annotation pipeline. Previously, BLASTP (9) and InterProScan (10) searches required up to 80% of the total CPU time of the PEDANT genome annotation workflow. To master the high number of newly sequenced genomes and to keep the data in PEDANT up-to-date, a radical reduction of this huge computational effort has become necessary.

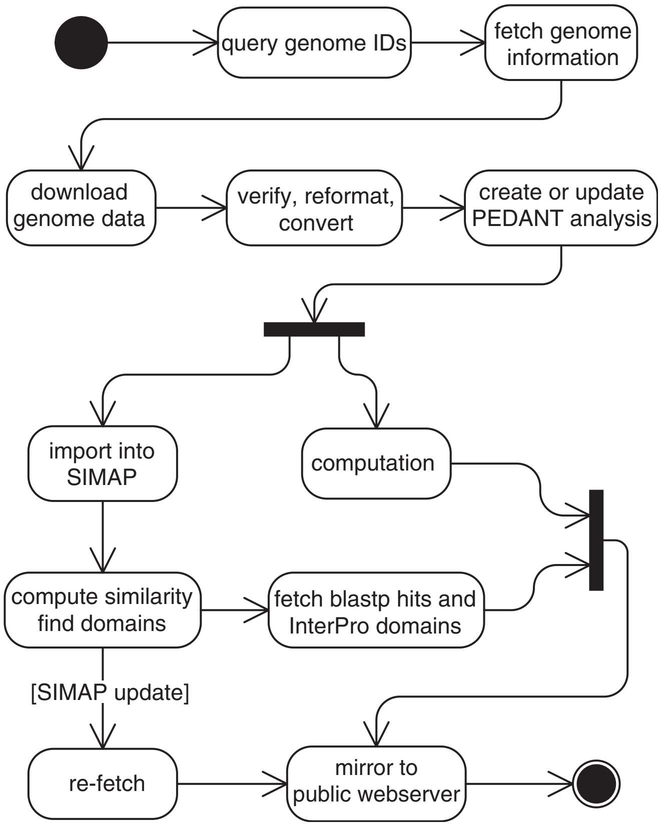
query genome IDs
fetch genome information
download genome data
verify, reformat, convert
create or update PEDANT analysis
import into SIMAP
computation
compute similarity find domains
fetch blastp hits and InterPro domains
[SIMAP update]
re-fetch
mirror to public webserver

Figure 1. UML activity model of the PEDANT genome import and processing pipeline. Symbols according to the UML 2.0 specification (http://www.uml.org) for activity diagrams.

The most obvious answer to this problem is to utilize high-performance computing facilities and avoid redundant calculations. The similarity matrix of proteins (SIMAP) (11) provides precalculated and up-to-date all-against-all alignments as well as domain assignments for essentially all publicly available protein sequences (21 million as of this writing). Our recent efforts to integrate PEDANT with SIMAP made it possible to avoid computationally intensive BLASTP and InterProScan runs and have led to a dramatic acceleration of the genome annotation work. Compared with *de novo* calculations, retrieving similarities and domains from the SIMAP database reduces the required CPU time by factors between 5 and 60. A typical bacterial genome with 3000 predicted genes can be processed at MIPS in <40 min using 60 Sun Grid Engine (SGE, http://gridengine.sunsource.net) nodes.

To generate and obtain these data, we have developed a computational workflow that coordinates the tasks between PEDANT and SIMAP. The first step in this workflow involves the import and maintenance of genome sequences and primary annotation provided by the respective source databases in PEDANT. In a subsequent step, SIMAP automatically retrieves protein and sequence data from PEDANT. If novel protein sequences previously unknown to SIMAP have been imported, their similarities to all other protein sequences and their domain architecture are calculated in SIMAP by utilizing large public resource computing facilities (12). As soon as the precalculated data are completely available in SIMAP, a notification event is triggered to start the SIMAP-based methods in PEDANT. These methods have been implemented as remote Enterprise Java Bean (EJB) invocations, which allow for rapid and efficient retrieval of data from SIMAP. One method designed to replace BLASTP retrieves homologs from a composite nonredundant database that includes PDB, UniProt/Swissprot, UniProt/TrEMBL, as well as all protein sequences already present in PEDANT. The second method which serves as a substitute for InterProScan retrieves precalculated protein domain assignments considering all InterPro member databases according to the InterPro XML format specification, except for the TMHMM (13), SignalP (14) and TargetP (15) methods which are run by PEDANT itself considering the appropriate genomic context (i.e. gram stain for signal peptides).

Web Services

The comprehensive collection of 3000 extensively annotated genomes provides a unique foundation for data mining and large-scale investigation of genome properties. While information on a limited number of genes of interest can be conveniently explored using the PEDANT web interface, any computational analysis of genomes at large necessitates local access to data. However, the large amount of annotation data computed for 4.5 million PEDANT proteins makes systematic dissemination of database dumps or flat files unpractical (although we do provide them upon request). Instead, we offer a simple, transparent and computer language-independent remote access based on the Web Service technology. This service has been implemented as a document style, SOAP-based Web Service (see http://www.w3.org/TR/soap12-part0). It can be easily integrated into own applications since for most computer languages libraries exist to access these kind of services. The functions provided by the Web Service are described in a Web Service Description File (WSDL, see http://www.w3.org/TR/wsdl), which allows for an automatic generation of a client program, e.g. by using the Perl SOAP::Lite (http://www.soaplite.com) or the Java Axis (http://ws.apache.org/axis/java/index.html) libraries.

The PEDANT3 WSDL File can be found at http://mips.gsf.de/webservice/pedant3/Pedant3AccessBeanService/Pedant3AccessWebService?wsdl. At present the service provides the following query types:

(i) return the list of organisms processed in PEDANT,
(ii) return the computational methods used to annotate a particular organism,
(iii) return a result overview (e.g. which functional category appears how many times) for a certain method in a certain organism,
(iv) return the genetic elements of an organism,
(v) return the result of a certain method for a single genetic element or for a whole genome ordered by its genetic elements.

For the latter query type it is possible to search in both directions: the service can return all genetic elements having a certain property (e.g. a certain functional attribute), or all properties of a certain genetic element (e.g. all functional attributes of a protein). Furthermore, in the former case it is possible to query several genomes at once. For BLASTP- and SIMAP-based methods, it is possible to restrict the results by an *E*-Value cutoff. A detailed overview of the Web Service functionality can be found at http://pedant.gsf.de/webservices.jsp.

The PEDANT3 Web Service encapsulates the complicated internal data structures of the PEDANT database and returns the results in a generic format that consists of key-value pairs of properties assigned to a given genetic element. This generic format assures that the end-user client software will not have to be reprogrammed if new methods are introduced into the PEDANT system.

DISCUSSION

There is no fixed release cycle for PEDANT. As soon as new genomes become available at RefSeq or any other listed genome resource, they will be imported, processed and made available via the web server. However, since SIMAP has a monthly release cycle, the computation of a genome by PEDANT is typically finished roughly 1 month after its import. Since the PEDANT3 software is now stable and all genomes from the previous version, PEDANT2, have been either migrated or reimported into PEDANT3, we took PEDANT2 and its Web Service offline. We also discarded all incomplete genomes previously available via PEDANT2 because the new high-throughput technologies now allow finishing genome sequencing projects on a very short-time frame.

In the future, genomes from further resources [i.e. USCS Genome Browser Database (16), Vega (17)] will be imported and previously imported genomes will be kept up-to-date. We are also in the process of supplementing the PEDANT web site by multiple new features, including viewing the genome project information [RefSeq status, source sequence centers, whole-genome shotgun (WGS) (18) sequencing coverage, number of records, etc.], taxonomic selection of genomes and improved search capabilities. A cross-genome index for precomputed annotations is nearly finished and will be available online shortly. This will allow for comparison of genomes based on their annotated features, such as domain content, functional categories and structural folds.

ACKNOWLEDGEMENTS

We are grateful to Volker Stümpflen for assistance with the Web Services.

FUNDING

Funding for open access charge: Helmholtz Gemeinschaft.

Conflict of interest statement. None declared.

REFERENCES

1. Frishman,D. and Mewes,H.-W. (1997) Pedantic genome analysis. *Trends Genet.*, **13**, 415–416.
2. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
3. Hubbard,T.J.P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.
4. Ruepp,A., Zollner,A., Maier,D., Albermann,K., Hani,J., Mokrejs,M., Tetko,I., Güldener,U., Mannhaupt,G., Münsterkötter,M. *et al.* (2004) The funcat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.*, **32**, 5539–5545.
5. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **36**, D13–D21.
6. Liolios,K., Mavromatis,K., Tavernarakis,N. and Kyrpides,N.C. (2008) The genomes on line database (gold) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, **36**, D475–D479.
7. Güldener,U., Mannhaupt,G., Münsterkötter,M., Haase,D., Oesterheld,M., Stümpflen,V., Mewes,H.-W. and Adam,G. (2006) Fgdb: a comprehensive fungal genome resource on the plant pathogen fusarium graminearum. *Nucleic Acids Res.*, **34**, D456–D458.
8. Kämper,J., Kahmann,R., Bölker,M., Ma,L.-J., Brefort,T., Saville,B.J., Banuett,F., Kronstad,J.W., Gold,S.E., Müller,O. *et al.* (2006) Insights from the genome of the biotrophic fungal plant pathogen ustilago maydis. *Nature*, **444**, 97–101.
9. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389–3402.
10. Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) Interproscan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.
11. Rattei,T., Tischler,P., Arnold,R., Hamberger,F., Krebs,J., Krumsiek,J., Wachinger,B., Stümpflen,V. and Mewes,H.-W. (2008) Simap–structuring the network of protein similarities. *Nucleic Acids Res.*, **36**, D289–D292.
12. Rattei,T., Walter,M., Arnold,R., Anderson,D. and Mewes,W. (2007) Using public resource computing and systematic pre-calculation for large scale sequence analysis. *Lect. Notes Bioinform.*, **4360**, 11–18.
13. Kahsay,R.Y., Gao,G. and Liao,L. (2005) An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes. *Bioinformatics*, **21**, 1853–1858.
14. Bendtsen,J.D., Nielsen,H., vonHeijne,G. and Brunak,S. (2004) Improved prediction of signal peptides: Signalp 3.0. *J. Mol. Biol.*, **340**, 783–795.
15. Emanuelsson,O., Nielsen,H., Brunak,S. and vonHeijne,G. (2000) Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
16. Karolchik,D., Kuhn,R.M., Baertsch,R., Barber,G.P., Clawson,H., Diekhans,M., Giardine,B., Harte,R.A., Hinrichs,A.S., Hsu,F. *et al.* (2008) The ucsc genome browser database: 2008 update. *Nucleic Acids Res.*, **36**, D773–D779.
17. Wilming,L.G., Gilbert,J.G.R., Howe,K., Trevanion,S., Hubbard,T. and Harrow,J.L. (2008) The vertebrate genome annotation (vega) database. *Nucleic Acids Res.*, **36**, D753–D760.
18. Staden,R. (1979) A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.*, **6**, 2601–2610.